Institute of Systems, Control and Information Engineers

システム/制御/情報, Vol. 46, No. 4, pp. 169–176, 2002

展　望

データマイニング展望

元田　浩*・鷲尾　隆*

1. はじめに

大容量記憶媒体の低価格化，計算機処理能力の向上，情報通信技術の急速な進展が，ネットワーク社会でのデータ収集・活用を容易にし，計算機可読な多くの情報がインターネットを通して世界中を飛び交い，これらの情報に誰もがいつでも自由にアクセスできるようになってきた．2001年3月時点でWebの総ページ数は40億頁に達しており，検索エンジンGoogleは8000台のPCを使いクラスタコンピューティングにより7000万件／日の検索を1件あたり0.5秒で処理している．

データマイニングはその名の示す通り，多量のデータから有用な知識を発掘する技術の総称であり，ここ数年，シーズとしての技術面からも，ニーズとしての応用面からも多くの注目を浴び，著しい進展を見せている．20年前のエキスパートシステムの到来，その後の過度の期待と失敗を思い出させるが，今回はそれとは事情が違い，確実に情報化社会の基盤技術として定着するように思われる．両者とも知識を全面に出している点では同じであるが，前者は人間の専門家の頭にある知識の抽出と利用を目的としており，後者はデータに内在する非明示的な知識を発掘し利用しようとする点が異なる．

本稿では，データマイニングにかかわる，1) 機械学習を中心とする最近の技術的進展を概観し，現状を整理し，今後の技術的な課題を指摘し，同時に，2) 我国を中心に現状でどのような分野にデータマイニングが使われはじめ，どのような問題点があり，今後どのような分野へと浸透していくかを予想し，将来の展望を述べてみたい．

2. 最近の技術の進展と課題

2.1 機械学習・データマイニングの最近の技術の進展

データマイニングは複数のプロセスからなり，少なくとも次の五つ[1]が含まれる [27]．1) 対象領域の理解，2) データの準備，3) パタン（知識）の発見，4) パタンの事後処理（視覚化，解釈など），5) 結果の活用．さらに，これらが繰り返されることが特徴的である．このうち，機械学習は主として3) のパタンの発見で使われている主要な技術，手法である．機械学習は端的にいえば「計算機が経験からどのくらい学べるか」に挑戦するものであり，この意味でもデータマイニングに必要不可欠な技術である．

機械学習も人工知能の研究の歴史と同じくらい長い歴史がある．多くの研究があり，いずれもデータに内蔵される一般的性質（パタン）の抽出（汎化）が目的ではあるが，大半はデータを分類するための規則や決定木を帰納するものである．また，データの形式は単純な属性－値のペアからなる表形式のものが大半である．評価の多くは複数の実データで有効であることを確認したといってはいるものの，これも大半はUCIのベンチマークデータに対するものであり，データの数，形式，属性の数，クラスの数，タスクの種類のいずれをとっても，データマイニングで扱おうとするものとは桁，質が違う．分類問題に限定しても，少数（といっても数万程度）のデータを念頭に考えられたアルゴリズムは原理的には大量のデータに対しても適用可能とはいえ，実際問題としては真に巨大なデータに対しては問題がある．

これらの状況に触発され，機械学習にもここ数年で大きな進展があった [6,23]．以下，これらの中から筆者らが主要な成果だと考えるものを簡単に概説する．各項目は必ずしも独立ではなく（直交する概念ではない），相互に関連するものもあることをご容赦願いたい．

アンサンブル学習　別名コミッティー学習ともいい，教師つき学習の精度向上に大きく寄与した．複数の分類器（アンサンブル）の結果を統合したものは個々の分類器の精度よりもずっと良くなるという性質を利用したものである．期待どおりの成果がでるためには個々の分類器の動作が互いに異なることが必要である．いま，分類器の数をL，個々の分類器の誤差をp（簡単のためすべての分類器で同じとする）とし，かつ誤差は独立（相関がない）であると仮定するとすると，L個中n個の分類器の結果が間違う確率は2項分布$B(p,L)$となるので，多数決の結果が間違う確率p'は$n=L/2$の仮説が間違う確率となり，$p<0.5$であれば$p' \ll p$となる．アンサンブルの作り方には色々ある．1) 訓練データのサブサンプリングを利用：復元抽出をするバギング（bagging) [2]，交差検定と同じ方法でデータを作る交差検定コミッティー

* 大阪大学　産業科学研究所

Key Words: data mining, machine learning, knowledge discovery, data mining application.

[1]参考文献 [8] では9個のプロセスに細分されている．

NII-Electronic Library Service

Institute of Systems, Control and Information Engineers

(cross validated committees)，間違ったデータの重みを増やすブースティング (boosting) [9] など，2) 入力の属性を変化：属性の部分集合を選択して訓練データを作成する方法（ニューラルネットでの適用例），3) 出力(目的) を変化：多クラス問題に対してクラスを二つのグループに分割し，分割の仕方を変えた複数の分割法を準備して結果を統合する error-correcting output coding と呼ばれる手法 [5]，4) ランダムネスを導入：決定木の各ノードで使用可能な属性集合をランダムに選択し，その中から最良のものを使う SAS[31]，決定木の各ノードの情報利得比の上位所定個数の候補からランダムに属性を選択する方法など，5) 階層の導入：複数の異なった学習アルゴリズムから得られた分類器で予測した結果を属性とするデータを作成し，これを別のアルゴリズムで分類する stacking[30]，ある学習アルゴリズムから得られた分類器で予測した結果を元の属性に追加し，それを別のアルゴリズムのデータとして分類する cascade など，多数の方法が提案されている．アンサンブル学習で，個別の分類器の精度はそれほど良くなくても，全体として予測精度を格段と向上させることが可能になった．しかし，データマイニングの観点からは結果の解釈が困難になり，理解しやすさの点で問題を呈している．たとえば100個の分類器の投票結果の解釈は絶望的である．

決定木，規則学習 決定木の学習には通常，トップダウンの分岐統括法が用いられる．アルゴリズムは簡単であるが，データ量が巨大になると，たとえばC4.5のようなアルゴリズムをそのまま適用すると処理効率が問題となる．大規模データへの対処法に関しての工夫もいくつかある．1) サンプリングの利用：分岐統括法では全データ数が少ないと，下のノードになるほどそこまでたどりつくデータが少なくなりフラグメント化が問題となるが，大規模データに対してはこれを逆手にとりルートノードに近いほど少数サンプルのデータを使って木を成長させることにより精度を全く落とさずに効率よく決定木を構築できる [26]．2) データ構造の工夫：決定木で一番時間のかかる部分は連続数値属性に対して，各ノードで各属性ごとに属性値の順にソートして情報利得比などの評価指標を求め最適な閾値を決定する部分である．SPRINT では各属性ごとに属性値でソートしたデータを別々のディスクファイルに分け，各ノードで属性が選択され閾値が決まる度にファイルを2分割（閾値より小と大）し，各データと分割されたファイルとの対応関係をハッシュ表で管理している．並列化も容易である [24]．SLIQ も類似のアルゴリズムを用いている．いずれの計算量もデータ数と属性数にほぼ線形であり数百万個のデータでも効率良く処理できる．3) アンサンブル学習：全データを N 個にランダムに分割し，各分割データ内で決定木を学習させ結果を投票する（交差検定コミッティー）[3]．並列化すれば N 倍高速化される．バギングとサンプリングを併用した方法 [29] などがある．4) 数値属性の離散化：事前に比較的少数個の区間に数値属性を離散化しておく方法である．多くの実用的な方法が提案されているが，大半は各属性を独立なものとして扱い，属性ごとに離散化している．

規則の学習には通常，被覆法（分離統括法）が用いられるが，決定木に対する工夫が基本的にはそのまま利用できる．多くのアルゴリズムが提案されているが，その中でも Ripper が効率的であることが知られている [4]．このアルゴリズムの計算量はデータ数 n に対して $O(n(\log n)^2)$ であり，C4.5RULE の $O(n^3)$ に比べて非常に高速化されている．

データ量の削減 精度を落とさずにデータ量を削減する方法は大きく二つの方法がある．一つは属性選択 (feature selection)・属性構築 (feature construction) [19,20]，他の一つは事例選択 (instance selection) [21] である．属性を選択すればそれだけ情報量が落ちるので精度は落ちると考えがちであるが，属性数が増えると属性空間内に学習に必要とされるデータ数が指数関数的に増加し，かえって学習が疎外される結果となる．したがって，不要な属性は最初から削除しておく必要がある．属性の数を n とすると属性の組合せの数は 2^n となるので，後で説明する相関規則の生成に必要な多頻度アイテム集合の探索と同じく上手な枝刈りをしないと計算量が天文学的に大きくなってしまう．また，学習にはバイアスが必要であり，学習アルゴリズム（知識表現も含めて）ごとにバイアスが違うため，厳密には学習アルゴリズムによって最適な属性の集合が変わってくる．このため評価指標として分類器の精度を用いるラッパー法の方がよいが，計算量の点で，通常は別の評価指標（クラスとの相関，不整合度など）を用いて事前に選択するフィルタ法が使われる．種々の探索制御ヒューリスティックスを使った手法が提案されているが，単純に一番効く属性から一つずつ追加していく方法や一番効かないものから一つずつ削除していく方法では，属性間に強い相関があるときによい結果が得られない．属性間の相関が扱えるものに近傍データでのクラス識別能力を予測して属性の重みを求める Relief ならびにその改良版がある [17]．しかし，扱える属性の数は数十どまりであり，データマイニングで扱う数千あるいは数万個の属性を対象としたものはまだないといってよい．筆者らの知る唯一の例外はオンライン線形閾値アルゴリズムを採用した Winnow である [18]．数万の属性を扱えるが，2クラスで，属性値は2値の問題に限定されている．一方，属性構築は，初期属性そのものを直接使うのではなく，初期属性のいくつかを組み合わせて新しい属性を定義して，それを使って問題を解決（分類器を学習）しようとするもので，組合せのための自由度が多く事前の領域知識に基づく指針なしにデータだけから的確な属性を構築することは至難の技である．決定木の各ノードで単一属性を用いたテストの代わりに，複数属性の線形関数を欲張り探索で求めたり，数個の限

NII-Electronic Library Service

Institute of Systems, Control and Information Engineers

られたオペレータの良好な組合せを遺伝的アルゴリズムで求めるなどの，限られたものしか実用化されていない．

事例選択に関しては上述したようにアンサンブル学習とサンプリングを組み合わせてデータ量を減らしてなおかつ精度を維持しようというものと，多量のデータから本質的なデータのみをうまく選択することにより，全データを使ったのと同じ精度を得ようとするものがある．後者の代表的な技術はサンプリングであるが，ほかにもいくつかある．サンプリングに関しては母集団からのランダムサンプリングが前提になるが，実際にはサンプルサイズと精度を理論的に評価することが難しい．許容誤差とその確信度を与えたサンプルサイズの上限を与える理論はあるが，非常に大きめなサンプルサイズを与えてしまい実用的でない．しかし，オンラインサンプリングに関しては，それまでに取り込んだサンプル数で十分かどうかを判定する逐次適応サンプリング（sequential adaptive sampling）に関するよい理論が得られており，その結果，精度を落とさずに二桁データ量を減らすことに成功しているとの報告もある [7]．バッチ処理であるが類似のものに漸進的サンプリング（progressive sampling）がある．サンプルサイズを徐々に増加させ，サイズと精度の関係を示す学習曲線がいつ飽和するかを予測し，サイズを決めるものである．ランダムサンプリング以外のものには，サポートベクトルマシンのサポートベクトルのようにクラス分布が変わる境界値付近のデータだけを選択するもの，決定木の各葉に到達するデータ数が均一になるように，各葉内のデータ数に逆比例してサンプルしその分データに重みをつけるもの，サンプリングを使わないものに類似のデータの平均値をとり代表点を生成（プロトタイプと呼ばれる）するもの（あるいはクラスタの中心で代用）などがある．目的がデータを特徴づけることなのか，他と差別化することなのかによって，削減のための方法が異なる．詳しくは参考文献 [21] を参照されたい．

領域知識の利用　データからの学習アルゴリズムの多くのものの弱点はすでに知っている知識をうまく活用できないことである．領域知識を一般的に扱うためには学習結果とデータとが同じ表現言語であればよい．データの一部として既存の知識を与えたり，学習結果を新たな知識として次の学習に活用できる利点がある．一階述語論理を表現言語とする帰納論理プログラム [10] はこれを可能にする学習法であり，逆伴意法によって正例を説明し負例を説明しない最弱仮説を求めるものである．精力的な研究が行われているが，膨大な探索空間を絞りこむための制御戦略がむずかしく，誰もが手軽に使えるほど洗練されてはいない．領域知識を利用する全く違ったアプローチにベイジアンネットに代表される確率ネットワークがある．ランダム変数間の確率的依存性をグラフ構造で表現し，確率分布をデータから推定（パラメータ学習）するものである．領域知識はグラフ構造に反映される．隠れ変数がない場合（たとえば教師つき学習）のパラメータはデータから簡単に求められる．隠れ変数がある場合（たとえばクラスの値が分からないクラスタリング）でも勾配法や EM 法（Expectation Maximization 法）で近似的に求められる．近年，非常に活発な研究が行われている分野であり，多くの応用例がある．構造そのものの学習法はまだ未完成であるが，いくつかの方法が提案されている．

ラベル無しデータの活用　ここ 2，3 年急速に注目を浴び始めた研究分野である．分類学習では分類結果（クラスラベル）の分かっているデータを使わなければならないが，ラベル付けには手間ひまがかかる．ラベルがついてないデータは多数あり，これを積極的に活用すべし，というものである．分類問題とクラスタリング問題の間を連続化して統一的に取り扱おうとする枠組みとも取れる．一例を紹介する [11]．二つの異なった学習アルゴリズムを用い，少数の同じデータでそれぞれ学習させる．各アルゴリズムには得手不得手があるので，それぞれがどのデータは自信を持って分類でき，どのデータはあまり自信がないかを判断できる．それぞれが自信があるラベルなしのデータを分類し，その結果を相手に教え，教えられた方は，それをラベルありデータに追加する．このようにして，相互に教育しあい，使えるデータを増やしながら精度を向上させる．この分野は，今後さらに発展するものと期待される．

相関規則　これは機械学習の分野で生まれた技術というよりはデータマイニング固有の技術といえるが，まとめて扱うことにする．相関規則を高速に求めるアルゴリズムとしては Apriori が著名である [1]．このアルゴリズムは最初からデータはメモリーに入りきらないくらい大きいことを想定しており，多頻度アイテムの数を k_{max} とするとディスクスキャン回数はたかだか $k_{max}+1$ 回でよいという特徴がある．その基本は多頻度アイテム集合の任意の部分集合は多頻度アイテム集合でなければならないという集合の包含関係に関する頻度（支持度）の単調性を用いた効率のよい枝刈りである．すなわち，要素数 k の多頻度アイテム集合の候補を，要素数 $k-1$ の多頻度アイテム集合のうち，$k-2$ 個の要素が共通で残り 1 個が互いに異なる二つの集合からのみ生成し，さらに，そのすべての $k-1$ 個の部分集合が多頻度アイテム集合であるもののみを真の候補とするアルゴリズムである．その後，AprioriTid，DIC，DHP など多くの改良が提案されて処理効率は徐々に向上したが，これらはいずれも Apriori 同様，多頻度アイテム集合を実際に生成している．要素数 100 の多頻度アイテム集合を求めるためには $2^{100} \simeq 10^{30}$ もの候補を生成しなければならないので，これらの方法では，どんな工夫をしても要素数が大きくなると限界がある．これに対し，最近提案された FP-Tree[12] は多頻度アイテム集合を生成しないで相関規則を求めることができる．しかもディスクスキャン回

NII-Electronic Library Service

Institute of Systems, Control and Information Engineers

数は2回でよい．1回目のスキャンで各アイテムの頻度を求め，各データ（トランザクション）を頻度順にソートする．2回目のスキャンで各トランザクション中のアイテムを接頭辞木（prefix tree）の形にマージし，共通に現われるアイテムをカウントしその数を記憶する．頻度順にソートしているので，接頭辞木の上部に行くほど各トランザクションに共通に現われるアイテムが来て，非常にコンパクトな木にすべての情報が記憶される．後は頻度の少ないアイテムからこの木を上にたどりながら，そのアイテムを含む多頻度アイテムの集合を数え上げていく．この方法でAprioriに対して一桁は処理効率が改善される．しかし，接頭辞木がメモリーに入りきらなければこの方法でもまだ問題がある．以上の手法は，いずれも相関規則としてはApriori同様，支持度と確信度を評価指標としており，FP-Tree以外は頻度の単調性を探索の枝刈りの根拠としている．したがって，FP-Tree以外は評価指標が単調性を有しないものは適用できないという大きな弱点を持つ[1]．よく知られているように確信度は相関規則の条件部と帰結部の実際の相関（correlation）を適正に表現していない．相関を表現するのにより適正なχ^2値などを使うためには，単調性の条件を緩和する必要がある．最近提案されたAprioriSMP [25]はこの問題に対して一つの解を与えるものである．この手法は凸性を有する評価関数（χ^2値，エントロピー，Gini指標など）に対して一般的に適用できる．すべての相関規則を求めることは諦めて，結論部を固定すれば，凸性を利用することにより条件部（アイテム集合）の上位集合の評価指標の上限を簡単に抑えることができることを利用し，評価指標を最大とする（あるいは上位指定個数の）条件部を効率よく求めることができる．Apriori同様，あるアイテム集合の上限値がそれまでの探索で得られた最大値（あるいは上位指定個数の値）より小さければ，上位集合を枝刈りできることを利用している．したがって，アイテム集合の候補の作り方もAprioriと同じ方法が採用できる．

構造データへの取り組み 相関規則に関しては多くの研究がありその進展は著しいが，アイテム同士には構造がない．時系列データの中から多頻度部分系列を発見する研究もあるが，より一般的な構造データ中の多頻度部分構造を効率的に発掘するアルゴリズムの研究は少ない．筆者らが開発したGBIやAGMはその数少ない研究の一つである．いずれもグラフ構造を対象とする．部分グラフ同型問題はNP困難であることが知られており，問題は基本的に難しい．GBIは連結するノードのペアを欲張り探索で逐次チャンクしながら多頻度部分構造を取り出す [22]．連結グラフしか扱えずかつ完全探索ではないが，グラフサイズにほぼ線形な計算量で多頻度部分グラフを取り出すことができる．AGMは（部分）グラフを隣接行列で表現し，Aprioriと同様の枝刈りをしながら部分グラフを完全探索する [15]．正規形，正準形の概念を導入した不要な候補を極力生成しないボトムアップアルゴリズムを採用し，効率よく多頻度部分グラフをすべて取り出すことができる．隣接行列を使っているので非連結グラフも取り扱うことができる．いずれも，発癌性有機性化合物など有害な化合物の同定や特徴的なWebのブラウジングパタンの発見に有力であるとの見通しを得ている．

その他 ほかにも，環境とインタラクションしながら効用期待値が最大になるようにアクションを決める一般的な方法を提供する強化学習，ノイズや異分子の検知，例外知識の発見，ニューラルネットワークからの規則抽出，テキストマイニング，ラフ集合を用いた知識発見，データからの法則発見など多くの分野で新しい進展があった．これらのいくつかは本特集号でも取り上げられている．

2.2 今後の課題

以上，機械学習を中心にデータマイニングに適用可能な主要技術の最近の進展を概観したが，大量のデータを扱うための工夫は見られるものの，基本的にはデータは1ヶ所に整理されており，数値や名辞属性で表現されたものしか扱えない．データマイニングの基になるデータは分散されており，かつその形式は数値や名辞属性だけの単純なものではなく，自然言語，画像（イメージ，地図）など多様であり，かつこれらが複雑に関係しあっている．これらを統合して扱える手法が必要となる．また，機械学習は自動化を第一目標としてきたが，データマイニングでは専門家との協調作業が必須であるので，専門家の介入の余地をもっと入れる必要がある．大規模問題での経験を積み，どこでアルゴリズムが破綻するかを見きわめることも大事であるが，全体を詳細にモデル化しようとする従来の試みは失敗する可能性がある．規範からのずれ，局所的な特徴抽出など，アルゴリズムのスケールアップ以上の質的変化が要求される．一方，機械学習の経験から凝った探索（数ステップの先読探索など）よりも簡単な探索（欲張り探索など）の方が往々にしてよい結果が得られること，柔軟性は不安定性の元となり，強力な表現が必ずしもよい結果を産むとは限らない（バイアスとバリアンスのトレードオフ）ことなど重要な知見も得られている．これらの知見を最大限活用し，サンプリング，統計量の保存などによるデータ量の削減（要約）やデータ圧縮技術を併用し可能な限りメモリー上で処理できるようにするなどの一層の工夫が必要である．さらに，データは全部使う必要があるか，あるいは全部使っても足りないかどうかを早い時期に推定する学習曲線の推定，傾きの統計的検定技術も重要となる．資源制約の中でいつ打ち切ってもよい効率的なオンラインアル

[1] FP-Treeは多頻度アイテムを効率よく取り出すことしかしていない．

NII-Electronic Library Service

Institute of Systems, Control and Information Engineers

第 1 表　各種分野におけるデータマイニングの代表的適用事例

金融分野
・マーケティング分野 潜在的な住宅ローン申込み顧客の推定 顧客に応じた銀行商品の適切な組合せ（クロスセールス）の設計提示支援 生命保険の潜在的解約候補顧客の発掘，効果的なダイレクトメール宛先候補顧客の発掘 ・業務特化分野 消費者ローン与信審査の半無人化ルールの発掘 顧客に応じたリスク細分型の自動車保険の設計提示支援，証券顧客と営業マンとのトラブル予測 社債格付け推測，クレジット・カードの不正利用パタン推定
流通・小売分野
薬局チェーン販売データからの優良顧客の発掘 投入時立上り売れ行きデータに基づく新製品販売予測 新製品のヒット要因分析，品物の売行き要因分析，牛乳販売量の予測 消費者購買行動パタンの分析，種々の販促条件下における併売パターンの分析
製造分野
ホームページでの顧客意見収集による次世代新製品開発（カスタマーリレーションマーケティング） 顧客の製品クレーム情報と製造情報の突合せによる設計・製造現場への品質管理要求発掘 製造現場の製造条件と製品検査結果の突合せによる製造工程の改善
通信分野
ホームページ閲覧情報からの個別顧客のプロファイリングと顧客傾向分析 電話回線網管理のための負荷状況把握や障害診断 電話網使用需要マーケティングのための通信トラフィックデータ分析 顧客の通話パタンによる通話回線不正使用検出 計算機システムへのアクセスログに基づく不正アクセス検出

ゴリズム（anytime algorithm）の開発も今後の課題である．

さらに重要なことは，機械学習は，現在は，属性選択などの前処理の一部を除き，マイニング（パタン発見）部分にしか使用されていない．冒頭に述べたようにデータマイニングは複数のプロセスを含み，その中でもデータの前処理に一番時間がかかる．人間が関与する割合が大きな部分であり，この部分の自動化に機械学習の適用が期待される．

一般的なことであるが，機械学習は分類を主としてきた．データマイニングのタスクはより広く，かつ機械学習としてのタスクが明確ではないものが多い．データマイニングのタスクをいかに機械学習の問題として定式化するかも重要な課題である．

3. データマイニング技術の応用の現状と展望

3.1 産業界におけるデータマイニングの現状と問題点

データマイニングに対する期待は大きく，技術的にはまだ大きな課題があるにもかかわらず，統計処理や決定木，ニューラルネットワーク，相関規則などの現状でも使える機械学習・データマイニング技術を使って，ここ 2，3 年の間に早くも産業界で実用的に使われ始めている．筆者らがサーベイしたものを第 1 表に示す．金融や流通などの分野におけるマーケティング調査への適用が主流であるが，この他にも製造業や通信サービス業における品質管理や顧客管理などへの適用も進んでいる．

公開された資料に基づく限り，最も多くの適用事例が見られるのは金融分野である．適用事例はマーケティング分野と各種金融業務に特化した分野に大別される．米国では，1994 年頃から流通業や金融業でデータマイニングの事例が報告されているが，我国でも近年は多くの事例が報告されるようになった [28]．この分野では，ニューラルネットワーク，コホーネンネット，クラスタリング，決定木，ラフ集合，重回帰分析など，多様なデータマイニング技術が用いられている．マーケティング分野では，膨大な顧客リストから候補を見いだす必要がある．この条件下で，生命保険の潜在的解約候補顧客や効果的なダイレクトメール宛先候補顧客のマイニングでは，業務の効率と質の改善効果が得られている．また業務特化分野では，与信審査半無人化ルールの適用による消費者ローン無人申込み機の開発や，膨大なクレジット・カード使用記録からの不正利用パタン発掘において実績を上げている．流通分野では，小売部門のマーケティングのためのデータマイニング適用が主流であり，POS データを用いた流通全般の業務知識の導出，インストアでの販売促進用知識の導出，有望顧客の洗出しなどが行われている [16]．データマイニング技術としては，決定木，バスケット分析，重回帰分析，相関解析などが用いられてい

NII-Electronic Library Service

Institute of Systems, Control and Information Engineers

る．売行き予測などに対する適用事例はまだ十分成功しているとはいえない．優良顧客の発掘や各種パタンの発掘・分析などで効果が上がっている事例が多いが，新しいデータマイニング技術よりも全体的傾向を把握する従来の統計的手法に依拠する事例が多い．これは金融分野に比べて扱う商品や小売条件，顧客行動パタンがはるかに多様であり，顧客や購買事例を把握容易な形で類別して特徴を発掘することが難しいためであると考えられる．製造分野におけるデータマイニングの適用は，他の分野と同様に進展を見せているが，多くは社内の文書やマニュアル検索，マーケティングへの適用であり，他の業種と似通った目的，技術適用となっている．その結果，現状では製造業固有のデータマイニング適用事例はあまり多く見あたらない．第1表に掲げたものは製造業固有の適用事例である．最初のものはカスタマーリレーションマーケティングへの適用であり，主要電機メーカ，家電メーカが試みているがまだ試行段階の域を出ていない．後二者のような品質や工程管理への適用は，今後広範な適用可能性を有しかつ現状においても実用化が進められている事例である[14]．このような適用事例では，事例ベース検索やテキストマイニング，バスケット分析，決定木などの最新のデータマイニング技術が用いられ，効果を上げている．通信分野では，おもにインターネットの顧客マーケティングや電話網管理の分野にデータマイニング技術が用いられている．使用技術は分類決定木，バスケット分析，ベイジアンネット，ニューラルネット，テキストマイニング，各種統計的手法など多岐にわたる．通信分野には豊富な電子化データ蓄積があるので，データマイニングの適用範囲は広い．特に不正使用や不正アクセスの検出など，膨大な通信ログから特徴的パタンを発掘する適用は成功を収めている[13]．

使用技術と実施体制の実像　以上に述べた適用事例を含め，産業界で用いられているデータマイニング技術は多様である．どのような技術が用いられるかは，各事例の目的やニーズ，データの仕様のみならず，データマイニングツールの開発者やユーザによっても左右される．特にわが国では製造，通信分野の開発者やユーザは技術的な蓄積を有するため，市販ツールを利用するのみならず，種々の技術をテストしその中から最良のものを選択して，適用対象にカスタマイズ，チューニングしたツールやシステムを自ら構築することが多い．これに対し金融や流通分野では，事例に適したツールをユーザ自らが開発することは少なく，製造や通信分野に比較すれば既成の市販ツールを用いてデータマイニングを実施することが多い．データマイニング技術の開発や使用時の体制も事例によって異なっている．製造や通信分野では，対象データ収集，技術開発やシステム開発，使用までをすべて自前で行う場合が多い．しかし金融や流通分野では，コンサルタントや開発企業と組んでシステム開発，ツール使用，あるいはスキーム開発を行ったり，市販ツールを購入しユーザとしてそのまま利用する場合が多い．

技術的および実務的問題点　以上の産業界における実像の中で，いくつかの問題点が浮かびあがってくる．その一つは「データ収集ボトルネック」である．現場に蓄積されたデータを利用する際には，データが特定のマイニングを目的として蓄積されたものではないため，必ずしも目的達成に必要な情報を含んでいないことがしばしばある．このような目的とデータ内容の不整合性は，実際にマイニング分析を実行しないと明らかにできないことも多い．低コストな補足データ収集手段の確保や，既存データからの必要情報の推定手法などの技術が重要となるが，現状ではまだまだ未開拓である．データマイニングの研究やその応用ではマイニングの本体技術ばかりが注目されるが，データ収集法の十分な検討・改良，補足処理など，データ収集技術こそが成功の鍵を握っているといっても過言ではない．他の一つはデータマイニングの実施体制において，往々にして市販ツールを購入してユーザが使用すれば何とかなるという「市販ツール万能主義」が採られることである．効果的なデータマイニングには，豊富な知識と経験に基づいて各事例に適した様々な技術の組合せや設定条件を見つける必要がある．マイニングの目的や対象データの内容は事例ごとに千差万別である．現状の市販ツールは個別技術やそれを組み合わせる環境を提供するものであって，事例に即した適切な技術の結合による処理スキームや各種性能評価指標，チューニングパラメータ設定までは教えてくれない．データマイニングの実務への適用においては，ユーザが十分な知識や経験を蓄積するための時間や資本を投入するか，コンサルタントや開発会社と密接な連携体制を敷くなどの投資が必要となる．

この問題に関連してデータマイニング技術を提供する研究開発者側にも「研究分野の分断問題」が横たわっている．データマイニング技術は，人工知能やデータベース，統計など複数分野の基礎技術に根ざすものであるが，これら個別分野の研究開発者の連携が必ずしもうまくいっているとはいえず，技術がばらばらに提供されている傾向がある．また，データマイニングの実施スキーム全体を通じた各種技術の組合せにおける整合性や，各種目的やデータ内容に根ざした処理スキームの体系化などに関する研究もまだまだ手付かずの状態である．ユーザの立場に立ち，必要となる補助技術，各種判断指標，体系的マイニングスキームの蓄積などに関する研究が必要である．

3.2 今後の展望

産業界におけるデータマイニング技術の利用は，成功事例が多数報告されるようになったとはいえ，まだまだ内容や人材が限定されており発展途上段階にある．今後とも継続的な技術進歩は続き，研究事例を経て現場に浸透していくであろう．現場の方では実経験の積み重ねに

NII-Electronic Library Service

Institute of Systems, Control and Information Engineers

より，データマイニング遂行に必要な様々な知識や経験の重要性の認識がさらに増すであろう．これに伴い，優秀なコンサルタントや開発企業の成長，ユーザや開発技術者への知識や経験の浸透，および市販ツールの機能拡張や改善が促され，産業界への適用は一層拡大すると思われる．電子商取引や知識管理への展開が進展するであろう．

適切なマイニングスキームの設計支援を行う技術開発や事例蓄積が実施され市販ツールに盛り込まれる一方，ユーザやコンサルタント，開発技術者が効果的なマイニングスキームを組む能力を養うことで，ツールと人材の両面から質の高いマイニングが可能となる．さらにデータマイニング研究分野の内容が深化，確立していくにつれ，データマイニング技術全体のあり方を念頭に研究開発を進める研究者や技術者も増加してくると予想される．そうなれば研究分野の分断問題も，長い目で見れば解消に向かうと考えられる．

データマイニングが産業界現場において真に有用な技術となることを当面妨げる可能性のある難題は，データ収集ボトルネックである．これはデータ欠如の問題であり，それを正攻法で埋めるためにはデータ収集コストの壁に直面する．データマイニング技術の研究者間でさえも，この問題の重要性が明確に認識されていない．しかし，補足データを収集する方法の整備や欠如情報を他のデータから推定する技術など，問題を軽減できる可能性も十分あり研究開発が待たれるところである．ブームに便乗するような誇大宣伝は謹むべきであり，現状技術の可能性と限界に関する正しい認識を持つことが必要である．

4: おわりに

本稿では機械学習技術との接点，産業界の現状を中心にデータマイニングの現状を概観し，課題と近未来を展望した．データマイニングは，ある目的の実現に向けて多様な技術を組み合わせる総合工学的な研究分野である．したがって，学術研究と産業界実践の距離が近い分野である．理論が重要なことは疑う余地はないが，現場のニーズを的確に拾い出し，それを解決するという明確な目的意識を持った地に足のついた研究が必要である．データ収集から発掘された知識の活用までの道筋は単純ではない．途中のプロセス間で専門家（ユーザ）が介在する多くのインタラクションが必要になる．データありきの一方通行的な受動的マイニングではなく，必要なデータの収集，ユーザの目的に則したマイニンング手法の選択，結果に対するユーザのフィードバックに対する迅速な対応と，これら3者の間のインタラクションの相乗効果がスパイラル的に昇華するアクティブマイニングを念頭においた取組みが重要となるであろう．

(2001年10月2日受付)

参考文献

[1] R. Agrawal and R. Srikant: Fast algorithms for mining association rules; *Proc. of the 20th Very Large Dada Base Conference*, pp. 487–499 (1994)

[2] L. Breiman: Bagging predictors; *Machine Learning*, Vol. 24, No. 2, pp. 123–140 (1996)

[3] P. K. Chan and S. J. Stolfo: Learning arbiter and combiner trees from partitioned data for scaling machine learninig; *Proc. of the 1st International Conference on Knowledge Discovery and Data Mining*, pp. 39–44 (1995)

[4] W. W. Cohen: Fast effective rule induction; *Proc. of the 12th International Conference on Machine Learning*, pp. 115–123 (1995)

[5] T. G. Dietterich and G. Bakiri: Solving multi-class learning problems via error-correcting output codes; *J. of Artificial Intelligence Research*, Vol. 2, pp. 263–286 (1995)

[6] T. G. Dietterich: Machine-learning research: four current directions; *AI Magazine*, Vol. 18, No. 4, pp. 97–136 (1997)

[7] C. Domingo, R. Gavaldà and O. Watanabe: Adaptive sampling methods for scaling up knowledge discovery algorithm; *in* [21], pp. 133–150 (2001)

[8] U. M. Fayyad, G. Piatesky-Shapiro and P. Smyth: From data mining to knowledge discovery: an overview; *Advances in Knowledge Discovery and Data Mining* (U. M. Fayyad, G. Piatesky-Shapiro, P. Smyth and R. Uthurusamy, (eds.)), pp. 1–34 (1996)

[9] Y. Freund and R. E. Shapire: Experiments with a new boosting algorithm; *Proc. of 13th International Conference on Machine Learning*, pp. 148–156 (1996)

[10] 古川，尾崎，植野：帰納論理プログラム，共立出版 (2001)

[11] S. Goldman and Y. Zhou: Enhancing supervised learning with unlabeled data; *Proc. of 17th International Conference on Machine Learning*, pp. 327–334 (2000)

[12] J. Han, J. Pei and Y. Yin: Mining frequent patterns without candidate generation; *Proc. of ACM SIGMOD International Conference on Management of Data*, pp. 1–12 (2000)

[13] K. Hashimoto, K. Matsumoto and N. Shiratori: Probabilistic modeling of alarm observation delay in network diagnosis; *Proc. of the 6th Pacific Rim International Conference on Artificial Intelligence*, pp. 734–744 (2000)

[14] 堀，落田，浜田，井村：電気製品の市場品質監視システム—データマイニング技術の応用；人工知能学会誌，Vol. 15, No. 5, pp. 813–820 (2000)

[15] 猪口，鷲尾，元田，熊澤，荒井：多頻度グラフパタンの完全な高速マイニング手法；人工知能学会誌，Vol. 15, No. 6, pp. 1052–1063 (2000)

[16] （株）日経リサーチ：POSデータに対するデータマイ

NII-Electronic Library Service

Institute of Systems, Control and Information Engineers

ニング事例集 (2000)
[17] I. Kononenko: Estimating attributes: Analysis and extensions of relief; *Proc. of the 7th European Conference on Machine Learning*, pp. 171–182 (1994)
[18] N. Littlestone: Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm; *Machine Learning*, Vol. 2, pp. 285–318 (1988)
[19] H. Liu and H. Motoda: *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic Publishers (1998)
[20] H. Liu and H. Motoda (Eds.): *Feature Extraction, Construction And Selection—Data Mining Perspective*, Kluwer Academic Publishers (1998)
[21] H. Liu and H. Motoda (Eds.): *Instance selection and construction for data mining*, Kluwer Academic Publishers (2001)
[22] 松田, 元田, 鷲尾:一般グラフ構造データに対するGraph-Based Inductionとその応用;人工知能学会誌, Vol. 16, No. 4, pp. 363–374 (2001)
[23] T. M. Mitchell: Machine learning and data mining; *Communications of the ACM*, Vol. 42, No. 11, pp. 31–36 (1999)
[24] J. Shafer, A. Agrawal and M. Mehta: SPRINT: A scalable parallel classifier for data mining; *Proc. of the 22nd International Conference on Very Large Databases*, pp. 544–555 (1996)
[25] S. Morishita and J. Sese: Traversing itemset lattices with statistical metric pruning; *Proc. of ACM SIGACT-SIGMOD-SIGART Symp. on Database Systems*, pp. 226–236 (2000)
[26] R. Musick, J. Catlett and S. Russell: Decision theoretic subsampling for induction on large database; *Proc. of the 10th International Conference on Machine Learning*, pp. 212–219 (1993)
[27] 元田, 鷲尾:機械学習とデータマイニング;人工知能学会誌, Vol. 12, No. 4, pp. 505–512 (1997)
[28] 小野:金融業におけるデータマイニングの応用;第18回日本SASユーザー会研究発表論文集, pp. 159–171 (1999)
[29] M. Terabe, T. Washio and H. Motoda: S^3bagging: Fast classifier induction method with subsampling and bagging; *Proc. of the 4th International Symposium on Intelligent Data Analysis*, pp. 177–186 (2001)
[30] K. M. Ting and I. H. Witten: Stacked generalization: When does it work?; *Proc. of the 15th International Joint Conference on Artificial Intelligence*, pp. 866–871 (1997)
[31] Z. Zheng and G. I. Webb: Stochastic attribute selection committees with multiple boosting: Learning more accurate and more stable classifier committees; *Proc. of the 3rd Pasific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 123–132 (1999)

著者略歴

元田 浩 (もとだ ひろし)

1943年3月24日生. 1967年3月東京大学大学院原子力工学専攻修士課程修了. 同年, (株) 日立製作所に入社. 同社中央研究所, 原子力研究所, エネルギー研究所, 基礎研究所を経て1995年退社. 現在, 大阪大学産業科学研究所教授 (知能システム科学研究部門, 高次推論研究分野). 原子力システムの設計, 運用, 制御に関する研究, 診断型エキスパート・システムの研究を経て, 現在は人工知能の基礎研究, とくに機械学習, 知識獲得, 知識発見, データマイニングなどの研究に従事. 工学博士. 人工知能学会, 情報処理学会, 日本ソフトウエア科学会, 日本認知科学会, AAAI, IEEE Computer Society, 各会員.

鷲尾 隆 (わしお たかし)

1960年10月30日生. 1988年3月東北大学大学院原子核工学専攻博士課程修了. 1988年から1990年にかけてマセチューセッツ工科大学原子炉研究所客員研究員. 1990年 (株) 三菱総合研究所入社, 1996年退社. 現在, 大阪大学産業科学研究所助教授 (知能システム科学研究部門). 原子力システムの異常診断手法に関する研究, 定性推論に関する研究を経て, 現在は人工知能の基礎研究, 特に科学的知識発見, データマイニングなどの研究に従事. 工学博士. AAAI, 人工知能学会, 計測自動制御学会, 情報処理学会, 日本ファジィ学会, 各会員.

NII-Electronic Library Service